THE ROLE OF PROBABILITY FOR UNDERSTANDING STATISTICAL INFERENCE

Manfred Borovcnik
University of Klagenfurt, Institute of Statistics, Klagenfurt
manfred.borovcnik@aau.at

Probability is the basis for intelligent actions and decisions in the face of uncertainty. That includes statistical inference as well as considerations of reliability, risk, and decision-making. Curricula have reduced approaches with respect to the nature of probability. With easy access to computer technology, simulation has become the predominant approach to teaching. Although simulation is an effective method to replace complicated mathematics, it reduces concepts to their frequentist part. This culminates in an approach to informal inference that makes probability and conditional probability redundant. However, the relevant properties of statistical inference require a comprehensive conception of probability to be shaped in the individual's cognitive system.

INTRODUCTION

Foundational papers that clarify the relationship between probability and statistical inference require a wide spectrum of probability interpretations. We show that a discussion of the advantages and drawbacks of different approaches is necessary. We need to pay more attention to the different meanings of probability (equally likely, frequentist, and subjectivist) and shape a solid understanding of conditional probability and the Bayes formula. To build on this idea, we outline five pillars that link probability to inference. The first two pillars draw on the rich experience of playing and analysing games to develop reliable intuitions. The third pillar covers subjectivist meaning. The fourth pillar focuses on the need to connect to statistical inference in the early phases of probability teaching, and the fifth pillar relates to the intersection between probability and risk. The paper advocates a multi-method approach to teaching (and learning) statistical inference that includes classical *and* Bayesian methods rather than a reduced framework. Curriculum reforms should focus on the key role of conditional probability for all approaches to statistical inference and on a sound understanding of probability within a wider interpretation of statistical inference.

CORNER PILLARS FOR LINKING PROBABILITY AND STATISTICAL INFERENCE

Difficulties in understanding statistical inference are legendary. Since the first attempts to revitalise probability teaching in the 1980's through inference, the didactic discussion has been dominated by proposals to simplify the methods. In his effort to describe probability literacy, Borovcnik (2021) lists corner pillars of probability that all are connected to statistical inference.

Beginning Early and Develop the Ideas and Concepts in a Spiral Way (No. 1)

We can introduce statistical inference at the earliest possible stage. Creative tasks that make one of the main purposes of probability visible right from the beginning date back to Varga (1983). With 9(!) year-old children, Varga investigated the behaviour of chance with respect to runs of heads and tails in coin tosses and gets into the middle of inferential considerations when the children are asked to judge whether a specific protocol of coin tossing was real or was "invented" by those who wrote the protocol. That makes probability relevant in games and allows enough time for the concept to emerge in children and for it to be revised and adapted to the purpose of decision-making in the form of a statistical test. As Fejes-Tóth et al. (2022) note, "familiarity with combinatorial thinking makes the introduction of statistical hypothesis testing feasible. … it is didactically helpful to base the introduction of a complex method for making decisions … on combinatorial skills" (p. 5).

Using Games Intelligently to Induce Sustainable Probabilistic Intuitions (No. 2)

Using games of chance intelligently, rather than routinely, is the striking counter argument to those who criticise games of chance as irrelevant to teaching. The coin toss experiment by Varga (1983) is evidence of intelligent use of games. Such games are also useful in shaping connections between two central meanings of probability, namely the classical interpretation of proportions and the frequentist meaning of probability. The classical meaning can be applied to cases with equally-likely outcomes. The frequentist meaning of probability is useful for general cases where an equally-likely argument

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

(possibly based on the physical symmetry of a device producing the elementary outcomes) is missing, but the idea of experiments repeated under the same conditions applies (as in physical experiments). Steinbring (1991) refers to a *complementarity* between the concepts of equally likely and frequency and notes that the conception of probability requires both aspects. The interplay between probability assumptions based on equally-likely cases and the development of relative frequencies in turn leads to issues of statistical inference. That is, to understand probability properly, questions of statistical inference play a key role. On the other hand, to develop an understanding of statistical inference, an adequate conception of probability is a prerequisite.

Build Thinking in Bayesian and Risk Terms as Early as Possible (No. 3)

Bayesian ideas relate to thinking—as opposed to applying routinely mathematical concepts—and to conditional probability, integrating an impact that may be attached to an event under uncertainty. Carranza and Kuzniak (2008) have highlighted the problematic nature of conditional probability and the Bayes formula. Problems arise from the failure to extend the modes of meaning of probability to a subjectivist interpretation, which is a qualitative evaluation of a statement in terms of an abstract weight index. Such an evaluation is often misunderstood in a sense as if it were feasible to assign an arbitrary value to the probability. A qualitative judgement about the probability of a statement, on the other hand, is based on a person's preference system towards statements and is a mathematical expression of preferences. The main difference with the other meanings is that probability here is a property of the person's preferences, whereas equally likely and a frequentist interpretation of probability are associated with an objective world (with the device that generates the outcomes or with the conditions of the physical process that leads to the outcomes of interest). The usual attributes associated with this are subjective and objective as if the person's judgement must be subjective (arbitrary) and the property of the process or device objective (scientific and undisputed, though perhaps unknown). Yet, we only recognise the person's judgement and the property of "the world." The person's judgement must be based on qualitative knowledge and is therefore anything but arbitrary. For more details, see Migon and Gamerman (1999) or the lively discussion in *The American Statistician*'s Teachers' Corner (Witmer et al., 1997).

These details include various meanings of probability and their justification by an axiomatic theory, which, incidentally, is provided by Kolmogorov (1956) for frequentist and by de Finetti (1937) for subjectivist probability. Bayes' formula is key to any procedure of statistical inference, which shows that statistical inference is void without a sound knowledge of conditional probability and a balanced conception of probability that includes the equally likely, frequentist, and subjectivist meanings of probability. To understand classical methods of statistical inference, it is necessary to develop a profound knowledge of probability, including Bayes' formula. Otherwise, the usual misinterpretations will occur. As Diepgen (1992) states, "The student nowadays can misinterpret the significance level in Bayesian terms only because he has nowhere learned about a Bayesian alternative to the significance test" (p. 52). This makes it clear that statistical inference requires a broader concept of probability that includes not only Laplace probability and the frequentist view, but also a subjectivist interpretation. Another advantage of Bayesian problems is that they naturally link probability with risk.

Linking Probability and Statistical Inference from Early Teaching (No. 4)

Apart from games of chance, where there is a clear notion of the magnitude of the probability of an event as a ratio of favourable to possible elementary outcomes, early probability teaching pays much attention to the empirical law of large numbers. This is meant to motivate a frequentist meaning of probability but immediately leads to the questions of what probability value is justified for an event under investigation. Moreover, how can one claim to have enough data to guarantee that an empirical estimate of probability is good enough? That means that a project is doomed to failure from the start if probability remains disconnected from statistical inference. This insight influenced curriculum efforts in the mid-1980's, and it soon became clear that statistical inference would widen the focus to include several probability interpretations.

It is a twist of the history of statistics education that in the attempt to expand the meaning of probability from the equally likely to a frequentist interpretation, it soon became clear that the methodology had to go beyond this and either integrate a subjectivist connotation or accept serious logical flaws (Hacking, 1965). This is precisely the dilemma that Carranza and Kuzniak (2008) have

identified in relation to Bayes' formula. It mirrors the controversy in the foundations of probability (1930's–1980's), where the case of inference (either Bayesian or statistical) became the source of dispute over which meaning was superior to probability (see Hacking 1965). In addition to debating which interpretation was appropriate for teaching probability, it became an urgent task to find ways of learning to reduce the complexity of statistical inference. Simulation and resampling methods were on the rise with the expansion of computer power in the 1990's. Yet, it was still only a vision until Cobb (2007) proposed replacing statistical inference entirely by resampling, not only for educational purposes but also for the discipline in general. Accordingly, an approach called *informal inference* developed, based on resampling. Rossman (2007) described informal inference as, "going beyond the data at hand and seeking to eliminate or quantify chance as an explanation for the observed data through a reasoned argument that employs no formal method" (p. 1). delMas (2017) called the approach *simulation-based inference*, but we use the terminology of informal inference (although we do not refer to the mass of other loose arguments subsumed under informal inference). Borovcnik (2021) summarises the criticisms of informal inference as follows: "[One criticism] was the reduction of probability to a degenerate frequentist conception; another was that statistical inference involves complex concepts such as type II errors and that this would no more be expressible within a pure resampling framework" (p. 3). Although informal inference reduces probability to its frequentist aspect and makes probability redundant because everything is solved by simulation, Batanero and Borovcnik (2016) focus on scenarios that are embedded in a context that naturally reduces the complexity and has an intuitively accessible meaning for the concepts involved. This approach corroborates the perception of quality indices of statistical tests as conditional probabilities, whereas in informal inference the character of such indices degenerates into absolute numbers if only it is possible to address them.

Develop the Twin Relation Between Probability and Risk (No. 5)

The twins of probability and risk emerged from a common historical development (see Borovcnik & Kapadia, 2018), which often makes it difficult to identify whether a trait or property belongs to one or the other. Situations under uncertainty are not only about quantifying the degree of uncertainty, probability, but also about the consequences of the outcomes, i.e., impact, which leads to the second key purpose of probability, namely risk. There is a vast literature on risk with inconsistent use of terms (see Borovcnik & Kapadia, 2018), which is very confusing. Yet, the problem is mainly whether risk should encompass the probability of an "adverse" outcome and its impact (cost, benefits), only the probability of the adverse outcome, or only the impact, or whether risk should not refer directly to the adverse outcome but indirectly to factors that can potentially cause the adverse outcome. Such risk factors are also called hazards. For a definition of risk, see Borovcnik (2015). For current considerations, one aspect of risk is striking because it blurs the perception of probability: it is difficult, if not impossible, to separate an assessment of probability from impact. If an impact is large, positive, or negative, people tend to neglect the probability of the event. Moreover, there is a difference in people's behaviour; according to Kahneman and Tversky (1979), they are risk-seeking in loss situations and risk-aversive in gain situations. Such psychological biases clearly show how difficult statistical inference is, apart from the mathematical details and methodological issues.

UNDERSTANDING STATISTICAL INFERENCE BY CONDITIONAL PROBABILITY

Conditional probability plays a key role in both Fisher's significance test and Neyman-Pearson's test policy. By analogy with medical diagnosis, the conditional probabilities "inverse" to Type I and II errors are crucial to the quality of a diagnosis of a particular person as opposed to a long-term quality index. Two situations are compared that have the same long-term errors but reflect a different quality of one-off decisions. This leads to the underlying prior probability of the disease and its essential role for inference, whether it is hidden (classical inference) or explicitly treated (Bayesian inference). Rather than favouring one approach, the methods of inference should be clarified by using conditional probabilities.

Fisher's Significance Test and Neyman and Pearson's Test Policy

In conceptualising inductive inference, Fisher (1971) focused on the so-called null hypothesis, H_0. He intended to provide an objective approach to inductive inference that avoided use of Bayes' formula (inverse probability), that is, instead of the probability of the hypothesis H given data *x,* i.e.,

P (H | *x*), he used the direct probability P (*x* | H). Fisher claimed that it is possible to infer causes from consequences and hypotheses from observations. He considered the distance between the data and the null hypothesis sufficient to reject the null if P ($x \mid H_0$) is small enough (for continuous distributions, probabilities must be replaced by likelihoods). According to Fisher, a significance test is a procedure for determining the probability of an outcome and more extreme outcomes given a null hypothesis with no effect or relationship. Regarding his preferred significance level of 0.05, Fisher emphasised that this did not mean that the researcher allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments with non-significant results. For Fisher, the significance level represented an *abstract measure of discrepancy* between data and the null hypothesis, so that the significance level of a test is void of any sampling interpretation. Neyman insisted that a test must consider an alternative in conjunction with the null and that a test represents a decision in which two different types of errors occur. He interpreted these errors as long-run frequencies. The next section presents the situation in the context of process control from which the ideas originated. Almost from the beginning, when they published their ideas on statistical testing, Fisher and Neyman were embroiled in controversy (Hubbard & Bayarri, 2003). What is used today as statistical inference is a hybrid that reconciles the major philosophical and methodological differences. We develop the idea that evaluating a statistical test for Type I and II errors misses the point.

Risk in Statistical Inference—Risk of Wrong Decisions in the Long Run

Rather than validating a null hypothesis in the face of empirical evidence, Neyman and Pearson developed their policy of repeated testing in the decision-oriented context of process control (Neyman & Pearson, 1928). For example, a decision is sought given data on two hypotheses: a normal level of defectives in production is 4%, and an unacceptable level is 10% defectives. A threshold must be set for the percent of defectives in the sample that allows the size of the decision errors to be "controlled" for the different scenarios. Because the context allows for repeated decision-making under the same conditions, it makes sense to interpret Type I and II errors in terms of relative frequencies *in the long run*. Yet, it should be noted that the error probabilities are not absolute probabilities at all, but *conditional* probabilities in relation to the respective scenario. A Type I error is the erroneous rejection of H_0, which is P("sample is in the rejection region" | H_0), and a Type II error is the erroneous non-rejection of H_0, which is P("sample is NOT in rejection region" | H_1). This insight that we are dealing with *conditional* rather than absolute probabilities is blurred by language that allows for descriptions that omit the restriction of the statements to the specific scenario, such as "given that," "conditional on," under the circumstance that the production is under control," etc.

Analogy Between Medicine and Statistical Tests—Risk of One-off Decisions in Medical Diagnostics

It is up to the reader to translate the above scenario into the context of diagnosing a particular disease when some laboratory parameters are used. If H_0 is subsumed by the state that the patient does not have the disease under investigation, and H_1, the patient does have it (Ca, for Carcinoma), then a Type II error means that the patient is diagnosed as disease-free, which is called negative (–), even though the patient has the disease. A diagnosis of positive (+) indicates that the person has the disease: Type I error = P (+ | H_0) and Type II error = P (– | H_1). In a radiologic clinic and screening, we may have data as in Table 1. In medical jargon, we speak of *specificity* for the probability of a negative diagnosis given H_0 (not this disease), and *sensitivity* for the probability of a positive diagnosis given H_1 (the disease is present); false positive and false negative correspond to Type I and II errors. The analogy between a statistical test and the decision associated with a diagnostic procedure makes it clear that there is a more relevant probability than the two types of errors for describing the quality of a diagnosis, namely the probability that a patient actually has the disease in question given that the person has a positive diagnosis, which is a *one-off decision*. This is called the *positive predictive value* (PPV). Analogously, the *negative predictive value* (NPV), is for the case of a negative diagnosis. We calculate these quality indicators—either per row or per column—that apply if we randomly select one person out of these groups. From the associated probabilities (in Table 2), the *quality indices used in statistical tests*, namely Type I and II errors (or sensitivity and specificity) *miss describing the key aspect of the quality of the diagnostic procedure*. The clinic and screening situations have the same indices. Yet, PPV and NPV are strongly dependent on the context, so that a positive diagnosis is valuable in the clinic although it is useless in screening.

Table 1. Patients in clinic and screening with their confirmed status of disease and the diagnosis

	Radiologic Clinic				Screening		
	–	+			–	+	
H_0 No	96 Specificity →	4 False pos. →	100	No	95232 NPV ↑	3968	99200
H_1 Ca	20 False neg. →	80 Sensitivity →	100	Ca	160	640 PPV ↑	800
	116	84	200		95392	4608	100000

Table 2. Quality indices or error of different types for the radiologic clinic and the screening data

	Prevalence	Sensitivity →	Specificity →	PPV ↑	NPV ↑
Clinic	50.0%	80.0%	96%	95.2%	82.8%
Screening	0.8%	80.0%	96%	13.9%	99.8%
Formula	P(Ca)	P(+ | Ca)	P(– | No Ca)	P (Ca |+)	P (No Ca | –)

CONCLUSION

The analogy between statistical tests and medical diagnosis helps to see that the prior probability of the null hypothesis is missing. The controversy in the foundations is due to the prior probability. Its status cannot be a frequentist probability, but is a qualitative degree of belief, a subjectivist probability. Thus, to simplify the complexity of inference in any didactically sensible way, we would get a caricature of the concepts involved. Any reduction of complexity based on a purely frequentist conception of probability fails to resolve the conceptual issues. Similarly, a Bayesian approach may be more intuitive and lead naturally to statistical inference (Albert 2002), yet it misses the full concept of probability. For this reason, Migon and Gamerman (1999) and Vancsó (2009) suggest teaching classical and Bayesian inference in parallel. Two statements by Vancsó's prospective teachers might convince readers, "I understood the confidence interval only after I had become more familiar with the Bayesian region of highest density" (p. 199) and "I really like the Bayesian method because I saw … why the people have different opinions ... Because different people may have different prior distributions" (p. 199).

Other problems for statistical inference are the case of small probabilities and the logic of repeated decisions. Borovcnik (2015) shows how difficult it is to obtain reliable information from data about a probability of only 10^{-4}, so that small probabilities have to be modelled using assumptions, leading to a *qualitative* rather than frequentist connotation. And, statistical inference is full of small probabilities. Furthermore, the optimal decision depends on whether a decision is made *one-off or repeatedly*. That means that an insurance company, for example, must pursue a different strategy than the person who takes out an insurance policy. In the medical field, a decision made by a state differs from an optimal decision made by a private individual because of the different logic of repeated decisions, not to mention the differences in the benefits and interests of stakeholders at different levels. Albert (2002) or Hoegh (2020) make an argument for including Bayesian ideas in the curriculum. Our considerations advocate a pluralistic perspective on probability that implies a comparative statistical inference, transferred from a philosophical analysis (Barnett, 1982) to the educational corner. This paper outlines its necessity.

REFERENCES

Albert, J. (2002). Teaching introductory statistics from a Bayesian perspective. In B. Phillips (Ed.), *Developing a statistically literate society. Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS6).* ISI/IASE. https://iase-web.org/documents/papers/icots6/3f1_albe.pdf?1402524960

Barnett, V. (1982). *Comparative statistical inference* (2nd ed.). Wiley.

Batanero, C., & Borovcnik, M. (2016). *Statistics and probability in high school*. Sense Publishers.

Borovcnik, M. (2015). Risk and decision making: The "logic" of probability. *The Mathematics Enthusiast, 12*(1–3), 113–139. https://doi.org/10.54870/1551-3440.1339

Borovcnik, M. (2021). Corner pillars of probability literacy. *Proceedings of the 63rd ISI World Statistics Congress*. International Statistical Institute. https://www.isi-web.org/files/docs/papers-and-abstracts/154-day3-ips078-corner-pillars-of-probability.pdf

Borovcnik, M., & Kapadia, R. (2018). Reasoning with risk: Teaching probability and risk as twin concepts. In C. Batanero & E. J. Chernoff (Eds.), *Teaching and learning stochastics* (pp. 3–22). Springer. https://doi.org/10.1007/978-3-319-72871-1_1

Carranza, P., & Kuzniak, A. (2008). Duality of probability and statistics teaching in French education. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference.* ICMI/IASE. https://www.stat.auckland.ac.nz/~iase/publications/rt08/T1P2_Carranza.pdf

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum. *Technology Innovations in Statistics Education, 1*(1). https://doi.org/10.5070/T511000028

delMas, R. (2017). *A 21st century approach towards statistical inference—Evaluating the effects of teaching randomization methods on students' conceptual understanding* [Paper presentation]. 61st International Statistical Institute World Statistics Congress, Marrakesh, Morocco.

Diepgen, R. (1992). Objektivistische oder subjektivistische Statistik? Zur Überfälligkeit einer Grundsatzdiskussion. *Stochastik in der Schule, 12*(3), 48–54.

Fejes-Tóth, P., Vancsó, Ö., & Borovcnik, M. (2022). Combinatorial thinking as key for introducing hypothesis testing—evaluation of the planned secondary-school reform in Hungary. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering and educating today's learners in statistics. Proceedings of the Eleventh Conference on Teaching Statistics (ICOTS 11)*. ISI/IASE.

Finetti, B. de (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales Institut Henri Poincaré, 7*(1), 1–68.

Fisher, R. A. (1971). *The design of experiments.* Oliver & Boyd. (Original work published 1935)

Hacking, I. (1965). *The logic of statistical inference*. Cambridge University Press.

Hoegh, A. (2020). Why Bayesian ideas should be introduced in the statistics curricula and how to do so. *Journal of Statistics Education, 28*(3), 222–228. https://doi.org/10.1080/10691898.2020.1841591

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p) versus errors (□) in classical statistical testing. *The American Statistician 57*(3), 171–182. https://doi.org/10.1198/0003130031856

Kahneman, D., & Tversky, A, (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (N. Morrison, Trans.; 2nd ed.). Chelsea. (Original work published 1933)

Migon, H. S., & Gamerman, D. (1999). *Statistical inference: An integrated approach*. Arnold.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I and II. *Biometrika, 20A*, 175–240; 263–294.

Rossman, A. (2007, August). *A statistician's view on the concept of inferential reasoning* [Paper presentation]. Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), Warwick, UK.

Steinbring, H. (1991). The theoretical nature of probability in the classroom. In R. Kapadia & M. Borovcnik, (Eds.), *Chance encounters* (pp. 135–167). Kluwer. https://doi.org/10.1007/978-94-011-3532-0_5

Vancsó, Ö. (2009). Parallel discussion of classical and Bayesian ways as an introduction to statistical inference. *International Electronic Journal of Mathematics Education 4*(3), 181–212. https://doi.org/10.29333/iejme/242

Varga, T. (1983). Statistics in the curriculum for everybody—How young children and how their teachers react. In D. R. Grey, P. Holmes, V. Barnett, & G. M Constable (Eds.), *Proceedings of the First International Conference on Teaching Statistics* (Vol. 1, pp. 71–80). Teaching Statistics Trust. https://iase-web.org/Conference_Proceedings.php?p=ICOTS_1_1982

Witmer, J., Short, T. H., Lindley, D. V. Freedman, D. A., & Scheaffer, R. L. (1997). Teacher's corner. Discussion of papers by D. A. Berry, J., Albert, & D. S. Moore. *The American Statistician, 51*(3), 262–274. https://doi.org/10.2307/2684895